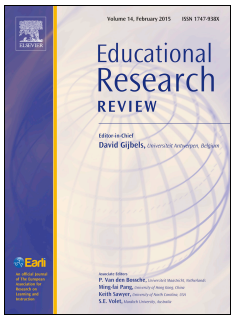
The use of Latin-square designs in educational and psychological research

Journal Item

Accepted Manuscript



The use of Latin-square designs in educational and psychological research

John T.E. Richardson

PII: S1747-938X(18)30162-3

DOI: 10.1016/j.edurev.2018.03.003

Reference: EDUREV 243

To appear in: *Educational Research Review*

Received Date: 12 June 2017

Revised Date: 12 March 2018

Accepted Date: 20 March 2018

Please cite this article as: Richardson, J.T.E., The use of Latin-square designs in educational and psychological research, *Educational Research Review* (2018), doi: 10.1016/j.edurev.2018.03.003.

The use of Latin-square designs in educational and psychological research

John T. E. Richardson

Institute of Educational Technology, The Open University. Walton Hall, Milton Keynes MK7 6AA, United Kingdom

E-mail address: John.T.E.Richardson@open.ac.uk

Acknowledgements

I am grateful to Carol Blumberg, Paul Ginns, Jimmie Leppink and Adrian Simpson for their comments on a previous version of this article.

The use of Latin-square designs in educational and psychological research

ABSTRACT

A Latin square is a matrix containing the same number of rows and columns. The cell entries are a sequence of symbols inserted in such a way that each symbol occurs only once in each row and only once in each column. Fisher (1925) proposed that Latin squares could be useful in experimental designs for controlling the effects of extraneous variables. He argued that a Latin square should be chosen at random from the set of possible Latin squares that would fit a research design and that the Latin-square design should be carried through into the data analysis. Psychological researchers have advanced our appreciation of Latin-square designs, but they have made only moderate use of them and have not heeded Fisher's prescriptions. Educational researchers have used them even less and are vulnerable to similar criticisms. Nevertheless, the judicious use of Latin-square designs is a powerful tool for experimental researchers.

Keywords: educational research; experimental design; Latin squares; psychological research

1. Introduction

A Latin square is a particular kind of configuration of integers, letters of the alphabet, or other symbols. Latin squares have been of interest to mathematicians for a very long time. However, Fisher (1925) proposed that they could also be very useful in experimental research for controlling the effects of extraneous variables. To be used properly, he argued that a Latin square needed to be chosen strictly at random from the universe of possible Latin squares that would fit a particular research design. He also insisted that the Latin-square design should be carried through into the analysis of the results.

Fisher was interested in agricultural experiments, but researchers in other fields came to realise that Latin-square designs could be useful in their work. This is notably the case in medical research, where it is nowadays widely recognised that Latin-square designs provide an efficient and effective way of controlling for the effects of extraneous variables, especially the effects of temporal order or sequence in repeated-measures designs. On February 7, 2017, the bibliographic database MEDLINE recorded a total of 4,055 publications since 1948 that contained the phrase “Latin square” in their titles, abstracts, keywords, or metadata, yielding an average of 58.8 such publications per year over the relevant 69-year period.

Educational researchers also realised that Latin squares could be used in their work, but Latin-square designs do not appear to have been adopted in education anywhere near as often as in medical research. Informal enquiries suggest that nowadays many educational researchers are unaware of their existence, and that their students do not learn about these designs in the course of their training. This is exceedingly unfortunate, because educational researchers may be missing the opportunity to exploit a potentially valuable tool in the design of their experiments. Accordingly, my aim in this article is to advocate the more widespread use of Latin-square designs in educational research.

To achieve this, I first review the history of Latin squares and their potential role in experimental research. I note that Latin-square designs are more efficient and hence more powerful than reasonable alternatives such as completely randomised designs or randomised complete block designs. I then review how educational researchers have made use of Latin-square designs in their experiments. In fact, such designs have been more widely adopted in psychological experiments, and so I also review how psychological researchers have made use of Latin-square designs. I compare research practice in these two disciplines with a focus on whether they have complied with Fisher's stipulations regarding the use of Latin-square designs in experiments. I conclude by advocating the more formal and rigorous use of Latin-square designs in future educational research.

2. What is a Latin square?

A Latin square is a grid or matrix containing the same number of rows and columns (k, say). The cell entries consist of a sequence of k symbols (for instance, the integers from 1 to k, or the first k letters of the alphabet) inserted in such a way that each symbol occurs only once in each row and only once in each column of the grid. Probably the best known modern examples are Sudoku puzzles, which will be discussed in Section 2.1. As a simpler example, Figure 1 shows a Latin square with four rows and four columns that contains the integers from 1 to 4. Figure 1 is an example of a standard form (also known as a reduced or normalised Latin square), in that the numbers in the first row and the numbers in the first column are in their natural order. Nonstandard Latin squares can be derived from standard forms by interchanging different rows in the grid, by interchanging different columns in the grid, or by doing both of these.

(Insert Figure 1 about here)

Latin squares are sometimes discussed in connection with magic squares. A (normal) magic square with k rows and k columns contains just one occurrence of each of the integers from 1 to k^2 in such a way that the numbers in each row, each column, and each diagonal add up to the same total (which simple mathematics shows must be $k(k^2 + 1)/2$). For example, a magic square with four rows and four columns would contain just one occurrence of each of the integers from 1 to 16, and the numbers in each row, column and main diagonal would all add up to $[4 \times (16 + 1)]/2$ or 34. (Non-normal magic squares can be constructed using more complex arithmetic progressions than 1, 2, . . . , k^2.) Latin squares and normal magic squares are conceptually different structures, but they are related in that Latin squares can be used to construct magic squares of the same dimensions (Emanouilidis, 2005).

2.1. A brief history of Latin squares

Kendall (1948) speculated that games and puzzles based upon Latin squares might have been first devised following the introduction of playing cards into Western Europe, which occurred during the 14th Century. In fact, Latin squares had been described by Arab and Hindu mathematicians prior to this and are depicted in early spiritual motifs found in the Middle East and India as well as Catalonia (Andersen, 2013).

Even so, the earliest written account in the West seems to be contained in a collection of puzzles and "recreations" by a French mathematician, Jacques Ozanam. The first edition had been published in two volumes in 1694; it described magic squares but not Latin squares. Ozanam died in 1718, but a "new edition, revised, corrected and augmented" was published posthumously as four volumes in 1723.[1] In a section entitled "Various Amusing Tricks", Ozanam (1723, p. 434) showed how to arrange the four kings, the four queens, the four jacks, and the four aces in a pack of cards in a 4×4 array so that there was a king, a queen, a jack,

and an ace in each of the four rows, in each of the four columns, and in both of the main diagonals, and so that there was a spade, a club, a heart, and a diamond in each of the four rows, in each of the four columns, and in both of the main diagonals. Figure 2 shows the solution presented by Ozanam (1723, Plate 12, Figure 35), but there are other solutions.

(Insert Figure 2 about here)

Ozanam's example goes beyond the simple notion of a Latin square in two respects. First, the solution involves not one but two Latin squares: one specifies the arrangement of the ranks (kings, queens, jacks and aces); the other specifies the arrangement of the suits (spades, clubs, hearts and diamonds). The two squares are orthogonal to each other, in the sense that each possible combination of ranks and suits only occurs once. Later, Euler (1782) discussed similar examples, using Latin (i.e., Roman) letters for the symbols in the first Latin square and Greek letters for the symbols in the second Latin square, and these became known as "Graeco-Latin" squares. Nowadays, however, they are more commonly referred to just as "pairs of orthogonal Latin squares" (or, occasionally, as "Eulerian squares").

Second, in Ozanam's solution, each symbol occurs only once in each of the main diagonals as well as in each of the rows and each of the columns. If the four ranks in the solution are represented as jacks = 1, aces = 2, kings = 3, and queens = 4, then the first of the two Latin squares can be represented as shown in Figure 3. It may be noted that the numbers 1, 2, 3 and 4 each occur once in each of the rows, in each of the columns, and in each of the main diagonals. This is known as a diagonal Latin square (Emanouilidis, 2005). This is not true of Figure 1, where the numbers 1, 2, 3, and 4 each occur once in each of the rows and in each of the columns but not in each of the main diagonals.

(Insert Figure 3 about here)

Since the beginning of the 20th Century, Latin squares have been studied in detail as interesting objects within the mathematical field of combinatorics (Andersen, 2013; Roberts

& Tesman, 2009; Wallis & George, 2011). The most authoritative account of Latin squares from a mathematical perspective is probably that by Keedwell and Dénes (2015). Nowadays, they are also encountered in daily life in the form of basic Sudoku puzzles: the solutions to these puzzles are 9×9 Latin squares, with the additional constraint that the integers from 1 to 9 each occur only once within each of the nine 3×3 squares that make up the overall 9×9 grid. (There are other varieties of Sudoku that adopt more complex constraints and some that use letters instead of integers.)

Birney, Halford and Andrews (2006) devised the "Latin Square Task" to measure the influence of relational processing complexity on human cognition. A participant is presented with a 4×4 Latin square in which some of the cell entries are hidden, together with the four symbols that are contained in the square. Their task is to say which symbol belongs in a nominated target cell so as to satisfy the requirements of a Latin square. Using data from both university students and school children, Birney et al. found that Rasch measurement analysis broadly confirmed their prior classification of different displays in terms of their complexity. Other researchers have confirmed the reliability and validity of this task as a measure of relational reasoning (Perret, Bailleux, & Dauvier, 2011; Zeuch, Holling, & Kuhn, 2011).

2.2. Latin squares in research design

Even so, this article is mainly concerned with the use of Latin squares in designing and implementing experimental research. One early example was a study conducted by a French agronomist, Cretté de Palluel (1788, 1790). He obtained four sheep of each of four different breeds and fed them on four different diets (potatoes, turnips, beets or corn). He drew up a schedule for slaughtering them over a period of four months which allowed for one

sheep in each of the four breeds and one sheep on each of the four diets to be slaughtered each month. The results showed the effect of the different diets while controlling for the breed of sheep and the month of slaughter. They enabled Cretté de Palluel to recommend that farmers should use root vegetables rather than the more expensive corn for fattening their animals during winter.

The use of Latin squares in experimental research in the modern era originated with the writings of Fisher (1925, pp. 229–232), who proposed that they could be used to arrange plots in agricultural experiments so as to control for differences in soil fertility:

> In a block of 25 plots arranged in 5 rows and 5 columns, to be used for testing 5 treatments, we can arrange that each treatment occurs once in each row, and also once in each column, while allowing free scope to change in the distribution subject to these restrictions. Then out of the 24 degrees of freedom, 4 will represent treatment; 8 representing soil differences between different rows or columns, may be eliminated; and 12 will remain for the estimation of error. (p. 229)

Fisher provided an example of this design in which mangel-wurzels (a variety of beet) had been planted in 25 plots and the weights of the yield were compared using an analysis of variance. The results are shown in Table 1. The variation across the five treatments was not statistically significant. (The treatment mean square is less than the residual mean square.) It might nevertheless be observed that together the variation among the rows and the variation among the columns accounted for 70% of the total variation in the yields of the 25 different plots. If the design had simply involved a comparison among 25 randomly chosen plots, the residual mean square would have been (7026.64 – 330.24)/20 = 334.82. The use of a Latin square meant that this was reduced to 146.19, rendering this arrangement a far more powerful

design for detecting differences among the treatments. Conversely, failing to incorporate the Latin-square design into the analysis of these data would have entailed a substantial loss of statistical precision and power.

(Insert Table 1 about here)

Fisher (1934) subsequently elaborated on the usefulness of this research design:

> In the Latin square any differences in fertility between entire rows, or between entire columns have been eliminated from the comparisons, and from the estimates of error, so that the real and apparent precision of the comparison is the same as if the experiment had been performed on land in which the entire rows, and also the entire columns, were of equal fertility. (p. 258)

In short, the removal of the variation among the rows and the variation among the columns leaves an unbiased estimate of the effect of the treatments, controlling for any differences among the rows and columns. Fisher went on to explain that the removal of extraneous sources of variation was a principle that could be "widely applied in all kinds of experimental work" (p. 258).

Fisher and Yates (1938) published a volume of statistical tables that might be used by researchers. They included the standard forms of all 4×4, 5×5, and 6×6 Latin squares, as well as examples of squares from 7×7 to 12×12, using letters of the alphabet to refer to the different treatments (pp. 44–46). To use these tables, Fisher and Yates stipulated that a standard form of the relevant size should be drawn at random from those in the published tables, that its rows should be interchanged at random using published tables of random sequences, that its columns should be interchanged in the same way, and finally that the

letters in the table should be assigned at random to the various treatments (p. 9).

The point of such randomisation was to avoid any systematic bias in the allocation of the *k* treatments to the rows and columns. Indeed, Fisher (1926) had previously argued that, in the absence of such randomisation, the resulting "systematic" arrangement did not strictly speaking count as a Latin square at all:

> The problem of the Latin Square, from which the name was borrowed, as formulated by Euler, consists in the enumeration of *every possible* arrangement, subject to the conditions that each row and each column shall contain one plot of each variety. Consequently, the term Latin Square should only be applied to a process of randomisation by which one is selected at random out of the total number of Latin Squares possible. (p. 510, italics in original)

Fisher was clearly overstating his point: a nonrandomised Latin square (such as that shown in Table 1) is still a Latin square. It might be more appropriate to suggest that the term "Latin-square design" should only be "applied to a process of randomisation by which one is selected at random out of the total number of Latin squares possible". Subsequently, Fisher (1937, pp. 78–99) elaborated this argument: "The process of randomisation, necessary to ensure the validity of the test of significance applied to the experiment, consists in choosing one at random out of the set of squares which can be generated from any chosen arrangement" (p. 80). In contrast, systematic arrangements which lacked the essential ingredient of randomisation were likely to lead to unreliable conclusions.

Using an example based on a 6×6 Latin square, Fisher (1937) also emphasised that the research design had to be carried through into the analysis of the results. In this case, the total of 35 degrees of freedom comprised 5 due to differences among the treatments, 5 due to

differences among the rows, and 5 due to differences among the columns, leaving 20 for the estimation of error. It was not appropriate to compare the variation among the treatments with the residual variation among the plots (i.e., with 30 degrees of freedom): the latter was likely to be inflated by the variation among the rows and the variation among the columns, which would thus undermine the precision and power of the experimental design (pp. 83–84).

In short, Fisher (1925, 1937) was making three points about Latin-square designs:

- The use of Latin-square designs provides a means of controlling the effects of extraneous sources of variation (the variables representing the rows and the columns).
- This can only be reliably achieved if the Latin-square design is chosen strictly at random from the universe of possible Latin squares that would fit the research design.
- The Latin-square design needs to be considered in the analysis of the results to achieve the increased statistical power that results from removing extraneous sources of variation.

Other research designs could of course be used to investigate such research questions, such as a completely randomised design or a randomised complete block design. The relative efficiency of different designs can be evaluated by comparing the error terms used to test the relevant effects. Yates (1935) reported that, in various agricultural trials carried out between 1927 and 1934, Latin-square designs had proved to be more efficient than other designs, and he argued that Latin-square designs could be useful in other fields of science and technology. Summarising Yates's results, Cochran (1938) stated that a Latin-square design had proved to be 2.22 times as efficient as a completely randomised design, whereas a randomised complete block design had proved to be just 1.67 times as efficient as a completely randomised design. This implies that a Latin-square design was 1.33 times as efficient as a randomised complete block design. Subsequent research confirmed that the Latin-square design was typically more

efficient and hence more powerful than reasonable alternatives (Kirk, 2013, pp. 688–689).

Nisbet (1939) was a teacher at the "demonstration school" set up to train teachers at the Edinburgh Provincial Training Centre. He published a study comparing four different ways of evaluating school children's spelling. On the basis of pre-testing, he assigned 100 words to four different lists of 25 words of approximately equal difficulty. He also assigned 80 pupils to four different groups of 20 pupils of equal spelling ability and tested the 80 pupils using the four different methods. The four tests were apparently administered in the same sequence to all four groups, but the four lists had been assigned as shown in Table 2. Nisbet noted that "each type of test involved all the words and all the pupils, although each word was given to each pupil only once" (p. 34). This eliminated any effects on the results of the four tests due to variations in the difficulty of the lists or in the ability of the pupils.

(Insert Table 2 about here)

Thomson (1941), who was the director of studies at the Training Centre, pointed out that, with one important exception, Nisbet's research design represented an application of the Latin square to an educational experiment. (Indeed, Nisbet might have used a Graeco-Latin square to manipulate the order of the four tests.) The exception was that the design in Table 2 had not been chosen at random but represented a particular systematic arrangement whereby the allocation of the four lists to the four groups in the four conditions was symmetrical around the major diagonal. More specifically, all four groups rotated through the same sequence of lists (A, B, C, D) with a different starting point. (The design is the same as the Latin square in Figure 1.) Following Fisher (1937), Thomson argued: "*In a true Latin square they* [the treatments] *should be arranged at random, subject, however, to the restriction that each treatment may occur only once in each row, and only once in each column*" (p. 135, italics in original). With this proviso, Thomson commended the use of Latin-square designs for future research involving educational experiments.

3. Latin squares in psychological research

3.1. Psychological discussions of Latin squares

Both Thomson and Nisbet were influential in educational research. The former held the Bell Chair in Education at the University of Edinburgh, while the latter was subsequently the first professor of education at the University of Glasgow. Nevertheless, Thomson was an educational psychologist, and it was psychologists rather than educational researchers who first became interested in the potential role of Latin squares in human research. Garrett and Zubin (1943) discussed the use of Latin and Graeco-Latin squares in psychological research in an article concerned with the analysis of variance, citing both Fisher (1937) and Thomson (1941) as sources. They argued that a need to control variations in spatial position was often important in experiments on psychophysics and spatial perception, while in other experiments there was a need to control variations in temporal order or sequence. These various contexts had in common a desire to counterbalance the order of administration of different conditions across different participants or groups of participants using a within-subjects design.

Garrett and Zubin described an unpublished experiment in which the participants had been asked to identify four colours at increasing levels of illumination. A 4×4 Latin square was used to assign the four different colours to four different levels of illumination in four different groups of participants, so that each participant only saw each colour at one level of illumination. However, Table 3 shows that the Latin square was constructed by rotating through the same sequence of colours: red, blue, yellow, and green. This is clearly a "systematic" arrangement and hence not a true Latin-square design in Fisher's (1937) terms. Nevertheless, Garrett and Zubin concluded that using designs based on Latin squares would enable researchers to eliminate extraneous sources of variation in their data analyses.

(Insert Table 3 about here)

Grant (1948) provided a detailed account of the use of Latin squares in the design and analysis of psychological experiments. Like a three-way factorial experiment, a Latin-square design contains three factors (row, columns, and treatments) that are statistically independent of one another. Unlike in a factorial experiment, the main effect of each of the three factors in a Latin-square design is confounded with the interaction between the other two factors. In particular, the treatment effect is confounded with the interaction between the effect of rows and the effect of columns. If the number of treatments is three or more, the confounding is only partial, and Grant stated that its expected value was zero if the Latin square had been selected strictly at random using the procedures described by Fisher and Yates (1938).

If only one participant is assigned to each row in the Latin square, Grant pointed out that it was not even possible to calculate the residual portion of the rows-by-columns interaction because it was needed as the error term in an analysis of variance (see Table 1). Whether or not it was appropriate for use as an error term could not be evaluated. For instance, using a within-subjects design such as that shown in Table 3, individual differences in the effects of practice might well inflate the rows-by-columns interaction. Grant argued that a solution to this would be to replicate each of the rows in the Latin-square design using several participants to estimate the subjects-by-columns interaction.

Bugelski (1949) noted that Latin-square designs control the ordinal position in which a treatment is administered, but they might not control the *sequence* in which treatments are administered. In Table 3, for instance, the colours red and blue each occur once at every ordinal position (first, second, third, and fourth), but red always precedes blue except when blue is presented first. Consequently, this design is highly vulnerable to carryover effects (if perceiving the colour red interferes with the ability to identify the colour blue, for example). Bugelski argued that in some fields of psychological research it was necessary or desirable to

control sequence or carryover effects as well as ordinal position (see also Edwards, 1951).

He noted that it is sometimes possible to control *immediate* sequence effects (that is, pairings between consecutive treatments) when the number of treatments is even. If each row in a Latin square represents the order of administration of the different treatments to different participants or groups of participants, Latin squares in which each consecutive pairing of two treatments occurs exactly once are described by mathematicians as "row-complete" Latin squares (see Keedwell & Dénes, 2015, pp. 70–72), although researchers sometimes describe them as "digram-balanced" Latin squares. Bugelski presented an example of a row-complete 6 × 6 Latin square, and Williams (1949) described a procedure for constructing row-complete Latin squares with any even number of rows and columns (see also Table 8 below).

However, row-complete Latin squares with odd numbers of rows and columns do not exist, and it is therefore not possible to control immediate sequence effects when the number of treatments is odd. A solution is to construct pairs of Latin squares that control sequence effects if used in combination (Tabachnick & Fidell, 2007, p. 514). Williams (1949), Bradley (1958), Wagenaar (1969), and Lewis (1989) all devised algorithms for constructing pairs of Latin squares that controlled immediate sequence effects, and Zeelenberg and Pecher (2015) described a procedure that controlled both immediate and more remote sequence effects.

Even so, there now arises a fundamental problem. Row-complete Latin squares are clearly systematic arrangements in Fisher's (1937) terms, not truly random Latin-square designs. Nevertheless, randomising the columns in a row-complete Latin square would be unlikely to lead to another row-complete Latin square. For instance, randomising the last five columns of a row-complete 6 × 6 Latin square leads to 120 possible Latin squares, but only four of these 120 Latin squares are themselves row-complete (Preece, 1991). In other words, researchers can randomise their choice of Latin squares or they can use Latin squares in order to achieve balance of carryover effects, but they cannot incorporate both in the same design.

Grant (1948) noted that one Latin-square design was especially problematic. This was the 2×2 Latin square, which has just one standard form and one nonstandard form, shown in Figure 4. In this design, the treatment effect is totally confounded with the rows-by-columns interaction. As Grant pointed out, it is therefore not possible to determine the treatment effect because of the possible presence of a rows-by-columns interaction. Even so, he also pointed out, this design is used frequently in experimental psychology when a researcher administers two conditions in a within-subjects design. Often, the two conditions are given in one order to half of the participants but in the reverse order to the other half. This "counterbalancing" does not control for interactions between the treatment variable and the counterbalanced variable.

(Insert Figure 4 about here)

Poulton and Freeman (1966) noted that 2×2 Latin-square designs were vulnerable to unwanted asymmetrical transfer effects. They recommended that researchers should always be prepared to carry out additional analyses to investigate the possibility of such effects in their research. Poulton (1982) suggested that asymmetrical transfer effects were the result of participants learning a strategy in one condition but employing it in a subsequent condition when it was unnecessary or inappropriate to do so. Poulton's specific proposals proved to be contentious, but there was broad agreement that the possibility of carryover effects does complicate the interpretation of results obtained using a 2×2 Latin-square design (Cotton, 1989). Indeed, such effects can arise in Latin-square designs using any number of treatments (Poulton & Edwards, 1979), and so the 2×2 design is not unique in this regard.

3.2. Psychological explanations of Latin squares

The apotheosis in psychologists' engagement with Latin-square designs occurred with

the publication in 1962 of Winer's *Statistical Principles in Experimental Design*, which contained a chapter of 64 pages on "Latin Squares and Related Designs" (pp. 514–577). Winer followed Fisher in emphasising the role of randomisation in selecting Latin squares:

> To obtain a Latin square for use in an experimental design, one of the standard squares of suitable dimension should be selected at random. The rows and columns of the selected square are then randomized independently. The levels of the factorial effects are then assigned at random to the rows, columns, and Latin letters of the square, respectively. (p. 517)

Winer went on to describe 13 different research designs or "plans" in which Latin squares might be used, of which nine involved repeated measures. The most complex design used a Graeco-Latin square to counterbalance the order of presentation of two different variables to different participants within a group but used the same Graeco-Latin square for each group. For each design, Winer explained the underlying statistical model (including the expected values of the mean square for each of the sources of variation) and provided the detailed hand calculations needed for the analysis of the data that might be obtained. For more complicated designs, Winer also discussed illustrative applications drawn from the published literature.

Winer published a second edition of his book in 1971 with relatively minor editorial changes to the chapter on Latin squares. He died in 1984, but a posthumous third edition was published in 1991 under the authorship of Winer, Brown, and Michels. This involved more substantive changes including the addition of new material and exercises at the end of the chapter, but they also removed particular sections, most notably the illustrative applications drawn from the literature. They summarised the potential uses of Latin squares as follows:

> Latin squares have essentially four related uses in research design in the social and behavioural sciences. They are typically used to control two or more nuisance variables, to counterbalance order effects in repeated-measures designs, to confound treatment conditions with group main effects, or as balanced fractional replications from a complete factorial design. (Winer, Brown, & Michels, 1991, p. 679)

Reese (1997) criticised this account because it tended to suggest that counterbalancing in itself always controlled the effects of the variable that was being counterbalanced and also that counterbalancing eliminated the effects of the counterbalanced variable from the effects of the treatments. He pointed out that, on the contrary, counterbalancing would only control the effects of the counterbalanced variable if it did not interact with the treatment effects. In addition, counterbalancing would only eliminate the effects of the counterbalanced variable from the treatment effects in very specific and highly unlikely circumstances.

Reese went on to suggest that many researchers had made the mistaken assumption that counterbalancing in itself would control for the effects of the variables that had been counterbalanced. As a result, they had not incorporated the effects of the counterbalanced variables in their data analyses and hence were unable to evaluate the effectiveness of their counterbalancing. However, Reese identified a more immediate reason for the researchers' failure to include counterbalanced variables in their analyses: that the statistical computer programs that were then available could not accommodate Latin-square designs. (As Reese pointed out, this is not an issue for designs using 2×2 Latin squares, where the main effect of treatments is actually identical to the interaction effect in a Rows $\times$ Columns analysis.)

This is not in fact an insurmountable problem. A solution is to carry out one analysis with a Rows $\times$ Columns design and another analysis with a Rows $\times$ Treatments design using

a statistical package of choice. The results can then be combined into a single analysis using the procedures that were described by Winer (1962, 1971). For instance, the research design devised by Nisbet (1939) (see Table 2) exemplifies Winer's "Plan 5". For this design, the first analysis would use the independent variables of groups and tests, and the second analysis would use the independent variables of groups and lists. The results could then be combined using the procedure described by Winer (1962, pp. 539–542). If it can be assumed that the interactions with the group factor are negligible, this yields complete information about the main effects of group, test and list, and partial information about the Tests × Lists interaction.

Analyses of this kind were described in detail by Tabachnick and Fidell (2001) in *Computer-Assisted Research Design and Analysis*, which included a chapter of 70 pages on "Latin-Square Designs" (pp. 481–550). They discussed several different research designs and provided worked examples using the command syntax in the computer packages SPSS, SAS, SYSTAT and MINITAB. They stressed the need to evaluate the distributional assumptions underlying these uses of analysis of variance and to provide measures of power and effect size. They also emphasised the importance of selecting Latin squares entirely at random and showed how to generate random Latin squares using the command syntax that was available in SAS and SYSTAT (pp. 522–524).

Tabachnick and Fidell (2007) published an updated version of this chapter in their subsequent book, *Experimental Designs Using ANOVA* (pp. 478–551). Resources using other kinds of software, such as the open-source programming language R, are also available (see, e.g., Todos Logos, 2010).

3.3. Psychological applications of Latin squares

The actual use of Latin squares by psychologists has remained modest but persistent.

The bibliographic database PsycINFO contains more than 4 million records, mainly relating to peer-reviewed publications. It subsumes the journal *Psychological Abstracts*, which went back to 1894, but it also contains some earlier publications. On February 7, 2017, PsycINFO recorded a total of 521 publications which contained the phrase “Latin square” in their titles, abstracts, keywords, or metadata dating between 1937 and 2016, yielding an average of 6.5 publications per year over the 80-year period. Nevertheless, many of their authors did not disclose how they had selected the Latin squares that they had used. As was noted in Section 2.2, Fisher (1926, 1937) had prescribed that Latin squares should be chosen at random, and this idea was endorsed by Thomson (1941) and Winer (1962, 1971). There is, in short, a clear possibility that many researchers over the last 80 years used systematic arrangements (as in the example given by Garrett & Zubin, 1943) rather than truly random Latin-square designs.

A simple Latin-square design of the sort shown in Table 1 can be analysed using a computer package by specifying rows, columns and treatments as the independent variables. Nevertheless, it remains the case that more complex Latin-square designs, especially those involving within-subjects variables, cannot be directly analysed using the statistical packages that are available today. (There is no syntax or option button to flag the use of a Latin-square design.) Indeed, the process cannot in principle be automated, because researchers need to decide for themselves in each specific case which terms in their analyses are confounded or assumed to be zero. As Reese (1997) suggested, this may well be a sufficient disincentive for researchers to incorporate into their data analyses variables that have been counterbalanced through the use of Latin-square designs.

Just as a snapshot, consider the eight publications identified by PsycINFO in the year 2016. Two (Daniel, 2016; Kuhn, 2016) described studies using the Latin Square Task that did not make use of a Latin square in their research design. One (Federer, 2016) was a doctoral dissertation that discussed the potential application of Latin squares to counterbalance the

administration of a number of assessment tasks. Three publications described studies in which 3×3 Latin squares had been used to counterbalance the order of administration of different conditions (Schippers, Schetters, De Vries, & Pattij, 2016; Yang, McClelland, & Furnham, 2016; Zack, Cho, Parlee, Jacobs, Li, Boileau, & Strafella, 2016), and two publications described studies in which 4×4 Latin squares had been used for this purpose (Agoglia, Holstein, Eastman, & Hodge, 2016; Willner-Reid, Whitaker, Epstein, Phillips, Pulaski, Preston, & Willner, 2016).[2]

In these last five publications, none of the authors disclosed how they had selected the Latin squares that they had used, which leaves open the possibility that they were systematic arrangements rather than truly random Latin-square designs as prescribed by Fisher (1926, 1937) and later by Winer (1962, 1971). The lead author of one of the papers reported that his team had devised a Latin square by rotating through the same sequence of three conditions (M. Zack, personal communication, January 30, 2017). As was noted earlier, this would have controlled the ordinal position of each condition, but it would not have controlled sequence effects. Moreover, none of the authors appears to have incorporated the Latin-square design of their studies in the analysis of the results. The same lead author reported that his team had specifically not included the sequence of treatments in their analysis because this would have reduced the degrees of freedom in error terms that were derived from a small sample ($N = 9$).

In short, Tabachnick and Fidell's (2007) book provides psychologists with the tools needed to make the best use of Latin-square designs. Nevertheless, only a modest number of psychologists appear to have availed themselves of Latin-square designs in experimental research, some have not taken care to choose their Latin squares entirely at random, and most have not taken account of their Latin-square designs in their data analyses. This suggests that they have not made use of Latin-square designs in the most effective manner. In particular, failing to incorporate the Latin-square design can lead to a major loss of statistical power.

As an illustration, consider the following hypothetical example. Six participants are asked to read three prose passages under three different instructional sets and to recall their content. The participants are assigned at random to three different groups, and a Latin square is used to assign the three prose passages to the different instructional sets (see Table 4). This exemplifies Winer's (1962, pp. 539–543) Plan 5. The participants' recall scores in the three instructional sets are shown in Table 5. Table 6 shows that an analysis of variance ignoring the Latin-square design yields a nonsignificant result, $F(2, 10) = 1.53$, $p = .26$.

(Insert Tables 4, 5, and 6 about here)

In contrast, Table 7 shows the results of an analysis of variance that incorporates the Latin-square design. (Only partial information is available concerning the interaction between the effect of instructional set and the effect of prose passage, which is why this term only has two degrees of freedom.) This analysis yields a highly significant effect of instructional set, $F(2, 6) = 32.17$, $p < .001$. There is also a substantial amount of variation across the recall scores associated with the three prose passages, $F(2, 6) = 101.54$, $p < .001$. In Table 6, the latter variation was pooled with the error term, which as a consequence was considerably inflated. As in Fisher's (1925) earlier example (see Section 2.2), taking the Latin-square design of the experiment into account radically reduces the residual mean square (in this case, from 28.02 to 1.33), yielding a far more powerful analysis of the same data set. Conversely, failing to take the Latin-square design of the experiment into account leads to a major loss of statistical precision and power.

(Insert Table 7 about here)

4. Latin squares in educational research

4.1. Educational discussions of Latin squares

As implied earlier in Section 3.1, educational researchers seem to have lagged behind their psychological counterparts in their appreciation of Latin squares. Lindquist (1940) had written a basic textbook, *Statistical Analysis in Educational Research*. However, according to Feldt (1979), by the 1950s Lindquist felt that this needed to be radically revised and updated to incorporate the advances that had been made in mathematical statistics in the intervening years. The result was an entirely new volume called *Design and Analysis of Experiments in Psychology and Education* (Lindquist, 1953). This included a brief account of the statistical principles underlying the use of Latin and Graeco-Latin squares (pp. 258–265), followed by an extended account of actual designs based on Latin squares (pp. 266–316).

Subsequently, however, relatively few educational researchers discussed the use of Latin-square designs. Houston (1967) discussed the problem of controlling sequence effects in repeated-measures Latin-square designs, as previously raised by Bugelski (1949). Houston followed Bradley (1958) in devising an algorithm for constructing pairs of Latin squares that controlled immediate sequence effects. He suggested that this would be useful in classroom experiments in which different teachers were assigned to teach a series of different classes. Houston did not consider how to control more remote sequence effects, but as noted earlier this was more recently addressed by Zeelenberg and Pecher (2015). Beall (1971, pp. 99–179) provided a more detailed account of the use of Latin squares in repeated-measures designs.

Collet and Maxey (1971) showed that data obtained using between-subjects designs involving Latin squares could be handled by means of multiple regression analysis rather than analysis of variance. It is indeed widely recognised that a between-subjects analysis of

variance can be treated as a special case of multiple regression (see, e.g., Pedhazur, 1997, pp. 4–5, 347–367, 405–414, 513–530). This would be the appropriate kind of analysis for agricultural experiments where the plants in different plots are independent of one another. Nevertheless, most uses of Latin squares in research with human beings involve within-subjects designs, and Collet and Maxey did not explain how their procedures could be adapted to this situation. In fact, most accounts of multiple regression analysis either ignore within-subjects (or longitudinal) designs completely (e.g., Pedhazur, 1997) or resort to using conventional analyses of variance (e.g., Cohen, Cohen, West, & Aitken, 2003, pp. 573–578).

Blumberg, Pearce, and Bader (1983) described an experiment on reading using adult participants which employed six different treatments that were applied to materials taken directly from six different texts in different subject areas. To meet the various constraints of the study, one 6×6 Latin square was used to assign the order of the six different treatments to six different groups, each of six participants. Each of the six participants in a group was given the same order of treatments. A second 6×6 Latin square was used to determine the order of administration of the six different texts to the six participants within each group.

This second Latin square was a row-complete Latin square that had been chosen specifically to control for immediate sequence effects (see Section 3.1): Table 8 shows that each of the six texts was followed only once in the Latin-square design by each of the other five texts. Blumberg et al. called this use of two different Latin squares in a non-standard manner a “matched Latin squares design”, and they explained how data obtained using this technique should be analysed. They concluded by commending this kind of design for use by other educational researchers. As was noted in Section 3.1, however, row-complete Latin-square designs are not truly random Latin-square designs in Fisher’s (1937) terms.

(Insert Table 8 about here)

4.2. Educational applications of Latin squares

The bibliographic data base of the Education Resources Information Center (ERIC) contains 1.5 million records of education-related materials. The collection was initiated in 1966, although it contains some earlier material. In the past, authors could deposit their own material, and so a proportion of the records relate to "grey" literature that has not been peer-reviewed. However, with effect from January 2016, ERIC introduced a selection policy that limited new records to material that has undergone some kind of review process. (This policy has not been applied retrospectively to the existing contents of the data base.)

On February 7, 2017, ERIC recorded a total of 69 documents which contained the expression "Latin square" in their titles, abstracts, keywords, or metadata originating between 1964 and 2016, yielding an average of 1.3 documents per year over the 53-year period, much less than the corresponding figure for PsycINFO. The 69 documents are listed in the online supplementary material for this article. Of the 69 documents, 34 are journal articles, 17 are conference presentations, nine are institutional reports, eight are doctoral dissertations, and one consists of other material that had been deposited with ERIC.

The 69 documents include eight articles concerning the use of Latin-square designs in educational research, four studies using the Latin Square Task that did not make use of a Latin square in their research design, nine articles by mathematics educators concerning the properties of Latin squares, and three articles about other uses of Latin squares. This leaves 45 documents reporting 44 empirical studies in which Latin squares had been used in their research design. Two studies used incomplete Latin squares (also known as Youden squares: see Tabachnick & Fidell, 2007, pp. 513–514). The other 42 used complete Latin squares: seven used 2×2 squares; 18 used 3×3 squares; nine used 4×4 squares; three used 5×5 squares; one used a 6×6 square; one used a 7×7 square; one used an 8×8 Latin square, one

used a 10×10 square; and one used a 16×16 square.

To determine whether there had been any variation in the interest (or lack of it) shown by educational researchers in Latin squares, the 44 studies were assigned to four groups based on their dates of publication. This showed that 16 studies had been published between 1964 and 1977, 15 had been published between 1978 and 1990 and 13 had been published between 2004 and 2016. This exercise did however reveal that ERIC includes no reports of empirical studies containing the phrase “Latin square” in their titles, abstracts, keywords, or metadata that were published between 1990 and 2003. Hamlin (2005) observed that a similar decline had occurred in marketing research in the 1970s, and he argued that marketing researchers had erroneously assumed that Latin squares were inferior to other kinds of research design. In contrast, in educational research there was a clear revival of interest during the 2000s, with no fewer than four empirical studies using Latin squares published in 2012 alone.

Strictly speaking, the choice of 2×2 Latin squares cannot be randomised, because there is only one standard form and one nonstandard form (see Figure 4 earlier). Although there is only one standard form of a 3×3 Latin square, there are 11 nonstandard forms, and so the choice of 3×3 Latin squares can be randomised. A fortiori, this is certainly true in the case of larger Latin squares. Ignoring the seven studies that had used 2×2 Latin squares, and ignoring the two studies that had used incomplete Latin squares, the authors of nine out of the 35 remaining studies did not indicate whether they had randomised their Latin squares.

In the other 26 studies, there was information in the reports of 18 cases to show that systematic arrangements rather than truly random Latin-square designs had been used. In each case, this was because the researchers had devised their designs by rotating through a single sequence of conditions. As was noted in Section 3.1, this would have controlled the ordinal position at which different treatments were administered but not the sequence in

which they were administered. As a result, these studies would have been highly vulnerable to carryover effects. Only in eight out of the 26 studies was there evidence that truly random Latin-square designs had been used. The point-biserial correlation coefficient between the year of publication of the 26 studies and whether or not a truly random Latin-square design had been used was -.09 (p = .68), suggesting that the failure to use truly random Latin-square designs has been a consistent feature of educational research over the last 50 years.

Another concern noted earlier is whether the use of a Latin square in an experimental design is carried through into the analysis of the data. The authors of only 17 of the 42 studies indicated that they had incorporated the Latin-square design in their analysis of their results. The point-biserial correlation coefficient between the year of publication of the 42 studies and whether or not the Latin-square design had been used in the data analysis was -.18 (p = .25), once again suggesting that the failure to carry through the Latin-square design into the data analysis has been a consistent feature of educational research over the last 50 years. As demonstrated by the hypothetical example presented in Section 3.3 earlier, this failure can undermine the potential statistical precision and power of a Latin-square design, because the resulting error term is inflated by the variation associated with the counterbalanced variable.

5. Conclusions

Like psychologists, educational researchers have had access to Winer's (1962, 1971) careful account of the construction and analysis of Latin-square designs, not to mention the earlier account by Lindquist (1953). More recently, they have had access to the explanations by Tabachnick and Fidell (2001, 2007) of how to analyse the results of experiments adopting such designs by means of modern computer packages. Readers are referred to these books for worked examples of studies using Latin-square designs. Winer and Tabachnick and Fidell

endorsed Fisher's (1926, 1937) original prescriptions that a Latin square needed to be chosen strictly at random from the universe of possible Latin squares that would fit a research design and that the Latin-square design should be carried through into the analysis of the results.

Fisher (1937, p. 80) argued that the use of systematic arrangements rather than a truly random Latin-square design would lead to unreliable conclusions. More specifically, Grant (1948) argued that the treatment effect would be confounded with the interaction between the effect of rows and the effect of columns in experiments that used systematic arrangements. Even so, around 70% of educational researchers who claimed to have adopted Latin-square designs over the last 50 years have used such arrangements. In particular, they have devised their Latin squares by rotating through a single sequence of conditions, which would have controlled the ordinal position of the treatments but not the sequence in which they were administered. As a result, their studies were highly vulnerable to carryover effects.

Fisher (1937, pp. 83–84) also argued that the Latin-square design needed to be carried through into the analysis of the results. In this situation, the Latin-square design is typically more efficient and hence more powerful than reasonable alternatives such as the completely randomised design or the randomised complete block design (see Cochran, 1938; Kirk, 2013, pp. 688–689; Yates, 1935). Ignoring the Latin-square design would mean that the variation among the rows and the variation among the columns (in other words, the variables being controlled) would be subsumed within the residual or error variation; this in turn would undermine the precision and power of the experimental design. Nevertheless, around 60% of educational researchers who claimed to have adopted Latin-square designs over the last 50 years had not made use of the Latin-square design in their analysis of their results. As the hypothetical example discussed in Section 3.3 makes clear, this is likely to be exceedingly unwise, because not incorporating the Latin-square design in one's analyses can lead to a major loss of statistical power (cf. Tabachnick & Fidell, 2001, p. 524; 2007, p. 524).

One limitation of the present critique of educational and psychological research is that it has relied on the bibliographic data bases ERIC and PsycINFO to identify studies that have made use of Latin squares. In some cases, these data bases do not contain the documents themselves, only their titles, abstracts, keywords and metadata, which leaves open the possibility that some relevant sources might have been missed. However, this would imply that certain educational or psychological researchers have been sufficiently innovative to make use of Latin squares in their experimental designs but did not see fit to mention this in their titles, abstracts, keywords, or metadata, which seems inherently unlikely.

To investigate this issue in more detail, all articles published in 2016 in *Learning and Instruction*, the sister journal of *Educational Research Review*, were examined. In total, 61 articles were published in six issues, each constituting a separate volume. Searching for the text "Latin" yielded a fair number of hits, mainly due to the use of words ending *-lating* or due to references to the ethnic category *Latino*. However, none of the 61 articles mentioned Latin squares or Latin-square designs. The implication is that such designs are not favoured by researchers whose work is published in *Learning and Instruction*. The present results can therefore be taken to give an accurate indication of the adoption or otherwise of Latin-square designs in educational and psychological research.

The mathematical properties of Latin squares are well understood (Keedwell & Dénes, 2015). For both educational and psychological researchers, the use of Latin squares in experimental designs provides a rigorous means of controlling extraneous sources of variation. They can be used to make more efficient use of limited resources and to achieve a higher level of statistical power than alternative experimental designs (Cochran, 1938). They are most likely to be used to counterbalance the order of administration of various conditions across different participants or groups of participants in a within-subjects design. This requires the use of special procedures to control both ordinal position and sequence effects

(Zeelenberg & Pecher, 2015), and these violate Fisher's (1937) requirement that Latin-square designs should be drawn entirely at random. It may also require the use of additional analyses to investigate the possibility of carryover effects (Poulton & Edwards, 1979).

Latin-square designs do have some disadvantages. For instance, the numbers of rows and columns need to be the same as the number of treatments (a requirement that is avoided by Youden squares and other incomplete designs). The number of participants should ideally be a multiple of the number of treatments (although cf. Willner-Reid et al., 2016). There may well be situations where within-subjects designs are simply not practicable because of subject attrition or carryover effects. More generally, it is inherent in Latin-square designs that some terms in the analysis are either confounded or assumed to be zero.

As Tabachnick and Fidell (2001, p. 486; 2007, p. 483) were at pains to emphasise, "The Latin-square arrangement, as an ANOVA [analysis of variance] model, assumes normality of sampling distributions, homogeneity of variance, independence of errors, and absence of outliers. Additional requirements for the repeated-measures application of Latin-square analysis are sphericity and additivity." Tabachnick and Fidell provided excellent advice on how researchers might assess these assumptions in their own experimental data using readily-available computer packages. Nevertheless, with these caveats, the judicious use of Latin-square designs is a powerful weapon in the armoury of experimental researchers, and it is for this reason that the aim of this article has been to advocate the more widespread use of Latin-square designs in educational research. Nowadays, educational researchers are being encouraged to use more sophisticated techniques (such as randomised control trials), and the increased use of Latin squares in the design and analysis of educational experiments would be compatible with this trend.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Agoglia, A. E., Holstein, S. E., Eastman, V. R., & Hodge, C. W. (2016). Cannabinoid CBI receptor inhibition blunts adolescent-typical increased binge alcohol and sucrose consumption in male C57BL/6J mice. *Pharmacology, Biochemistry and Behavior, 143*, 11–17. doi:10.1016/j.pbb.2016.01.009

Andersen, L. D. (2013). Latin squares. In R. Wilson & J. J. Watkins (Eds.), *Combinatorics: Ancient and modern* (pp. 251–284). Oxford, UK: Oxford University Press.

Beall, G. (1971). *Change-over experiments in practice*. Princeton, NJ: Educational Testing Service. Retrieved from ERIC database. (ED056062)

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: The development of the Latin Square Task. *Educational and Psychological Measurement, 66*, 146–171. doi:10.1177/0013164405278570

Blumberg, C. J., Pearce, D. L., & Bader, L. A. (1983, April). *Matched Latin squares: A nifty solution to a tricky design problem*. Paper presented at the 67th Annual Meeting of the American Educational Research Association, Montreal, Canada. Retrieved from ERIC database. (ED233058)

Bradley, J. V. (1958). Complete counterbalancing of immediate sequential effects in a Latin square design. *Journal of the American Statistical Association, 53*, 525–528. doi:10.2307/2281872

Bugelski, B. R. (1949). A note on Grant's discussion of the Latin square principle in the design of experiments. *Psychological Bulletin, 46*, 49–50. doi:10.1037/h0057826

Cochran, W. G. (1938). Recent work on the analysis of variance. *Journal of the Royal Statistical Society, 101*, 434–449. doi:10.2307/2980213

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple*

regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Collet, L. S., & Maxey, J. H. (1971). Analysis of variance and Latin Square problems by multiple regression analysis. *Journal of Experimental Education, 39*(4), 26–30. Retrieved from ProQuest database. (ProQuest document ID 1299995101)

Cotton, J. W. (1989). Interpreting data from two-period crossover design (also termed the replicated 2 × 2 Latin square design. *Psychological Bulletin, 106*, 503–515. doi:10.1037/0033-2909.106.3.503

Cretté de Palluel, F. (1788). Sur les avantages et l'économie que procurent les racines employées à l'engrais des moutons à l'étable [On the advantages and economy of using root vegetables to feed farm sheep]. *Mémoires d'Agriculture, d'Économie Rurale et Domestique,* Trimestre d'Été, 17–23. Retrieved from http://gallica.bnf.fr/ark:/12148/bpt6k5849689n/f60

Cretté de Palluel, F. (1790). On the advantage and economy of feeding sheep in the house with roots. *Annals of Agriculture, 14*, 133–139. Retrieved from https://books.google.co.uk/books?id=7JwZAQAAIAAJ

Daniel, E. (2016). Motivational and cognitive correlates of avoidance of ambiguity: The role of values and relational complexity. *Personality and Individual Differences, 102*, 149–152. doi:10.1016/j.paid.2016.07.001

Edwards, A. L. (1951). Balanced Latin-square designs in psychological research. *American Journal of Psychology*, *64*, 598–603. doi:10.2307/1418200

Emanouilidis, E. (2005). Latin and magic squares. *International Journal of Mathematical Education in Science and Technology, 36*, 546–549. doi:10.1080/00207390412331336201

Euler, L. (1782). Recherches sur une nouvelle espèce de quarrés magiques. *Verhandelingen*

uitgegeven door het zeeuwsch Genootschap der Wetenschappen te Vlissingen, 9, 85–239.

Federer, M. R. (2016). *Investigating assessment bias for constructed response explanation tasks: Implications for evaluating performance expectations for scientific practice* (Doctoral dissertation). Ohio State University, Columbus, OH. Retrieved from University Microforms International (Order Number AAI3731329)

Feldt, L. S. (1979). Everet F. Lindquist 1901–1978: A retrospective review of his contributions to educational research. *Journal of Educational Statistics, 4*, 4–13. doi:10.2307/1165062. Retrieved from http://www.jstor.org/stable/1165062

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd. Retrieved from http://psychclassics.yorku.ca/Fisher/Methods/

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 503–513. Retrieved from http://hdl.handle.net/2440/15191

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. London, UK: Oliver & Boyd.

Garrett, H. E., & Zubin, J. (1943). The analysis of variance in psychological research. *Psychological Bulletin, 40*, 233–267. doi:10.1037/h0063637

Grant, D. A. (1948). The latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin, 45*, 427–442. doi:10.1037/h0053912

Hamlin, R. P. (2005). The rise and fall of the Latin Square in marketing: A cautionary tale.

European Journal of Marketing, 39, 328–350. doi:10.1108/03090560510581809

Houston, T. R., Jr. (1967). *On the construction of Latin squares counterbalanced for immediate sequential effects* (Technical Report No. 25). Madison, WI: University of Wisconsin, Wisconsin Research and Development Center for Cognitive Learning. Retrieved from ERIC database. (ED013981)

Keedwell, A. D., & Dénes, J. (2015). *Latin squares and their applications* (2nd ed.). Amsterdam: North-Holland.

Kendall, M. G. (1948). Who discovered the Latin square? *The American Statistician*, *2*, 13. doi:10.2307/2682684. Retrieved from http://www.jstor.org/stable/2682684

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Kuhn, J.-T. (2016). Controlled attention and storage: An investigation of the relationship between working memory, short-term memory, scope of attention, and intelligence in children. *Learning and Individual Differences*, *52*, 167–177. doi:10.1016/j.lindif.2015.04.009

Lewis, J. R. (1989). Pairs of Latin squares to counterbalance sequential effects and pairing of conditions and stimuli. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1223–1227). Santa Monica, CA: Human Factors Society. doi:10.1177/154193128903301812

Lindquist, E. F. (1940). *Statistical analysis in educational research*. Boston, MA: Houghton Mifflin.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston, MA: Houghton Mifflin.

Nisbet, S. D. (1939). Non-dictated spelling tests. *British Journal of Educational Psychology*, *11*, 29–44. doi:10.1111/j.2044-8279.1939.tb03191.x

Ozanam, J. (1723). *Récréations mathématiques et physiques* [Mathematical and Physical Recreations] (new ed., Vol 4). Paris, France: Jombert. Retrieved from http://reader.digitale-sammlungen.de/de/fs1/object/display/bsb10594195_00005.html. Plates published separately. Retrieved from http://reader.digitale-sammlungen.de/de/fs1/object/display/bsb10594196_00005.html

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Perret, P., Bailleux, C., & Dauvier, B. (2011). The influence of relational complexity and strategy selection on children's reasoning in the Latin Square Task. *Cognitive Development*, *26*, 127–141. doi:10.1016/j.cogdev.2010.12.003

Poulton, E. C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin, 91,* 673–690. doi:10.1037/0033-2909.91.3.673

Poulton, E. C., & Edwards, R. S. (1979). Asymmetric transfer in within-subjects experiments on stress interactions. *Ergonomics*, *22*, 945–961. doi:10.1080/00140137908924669

Poulton, E. C., & Freeman, P. R. (1966). Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin, 66*, 1–8. doi:10.1037/h0023427

Preece, D. A. (1991). Latin squares as experimental designs. In J. Dénes & A. D. Keedwell (Eds.), *Latin squares: New developments in the theory and applications* (*Annals of discrete mathematics*, No. 46, pp. 317–342). Amsterdam: North-Holland.

Reese, H. W. (1997). Counterbalancing and other uses of repeated-measures Latin-square designs: Analyses and interpretations. *Journal of Experimental Child Psychology*, *64*, 137–158. doi:10.1006/jecp.1996.2333

Roberts, F. S., & Tesman, B. (2009). *Applied combinatorics* (2nd ed.). Boca Raton, FL: CRC

Press.

Schippers, M. C., Schetters, D., De Vries, T. J., & Pattij, T. (2016). Differential effects of the pharmacological stressor yohimbine on impulsive decision making and response inhibition. *Psychopharmacology, 233*, 2775–2785. doi:10.1007/s00213-016-4337-3

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis*. Boston: Allyn & Bacon.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Belmont, CA: Thomson Brooks/Cole.

Thomson, G. H. (1941). The use of the Latin Square in designing educational experiments. *British Journal of Educational Psychology, 11*, 135–137. doi:10.1111/j.2044-8279.1941.tb02156.x

Todos Logos. (2010, January 6). Latin squares design in R [Web log post]. Retrieved from https://www.r-bloggers.com/latin-squares-design-in-r/

Wagenaar, W. A. (1969). Note on the construction of digram-balanced Latin squares. *Psychological Bulletin, 112*, 881–911. doi:10.1037/h0028329

Wallis, W. D., & George, J. C. (2011). *Introduction to combinatorics*. Boca Raton, FL: CRC Press.

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research, Series A: Physical Sciences, 2*, 149–168. doi:10.1071/CH9490149

Willner-Reid, J., Whitaker, D., Epstein, D. H., Phillips, K. A., Pulaski, A. R., Preston, K. L., & Willner, P. (2016). Cognitive-behavioural therapy for heroin and cocaine use: Ecological momentary assessment of homework simplification and compliance. *Psychology and Psychotherapy, 89*, 276–293. doi:10.1111/papt.12080

Winer, B. J. (1962). *Statistical principles in experimental design*. New York, NY: McGraw-

Hill.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York, NY: McGraw-Hill.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Yang, J., McClelland, A., & Furnham, A. (2016). The effect of background music on the cognitive performance of musicians: A pilot study. *Psychology of Music*, *44*, 1202–1208. doi:10.1177/0305735615592265

Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, *2*, 181–247. doi:10.2307/2983638

Zack, M., Cho, S. S., Parlee, J., Jacobs, M., Li, C., Boileau, I., & Strafella, A. (2016). Effects of high frequency repeated transcranial magnetic stimulation and continuous theta burst stimulation on gambling reinforcement, delay discounting, and Stroop interference in men with pathological gambling. *Brain Stimulation*, *9*, 867–875. doi:10.1016/j.brs.2016.06.003

Zeelenberg, R., & Pecher, D. (2015). A method for simultaneously counterbalancing condition order and assignment of stimulus materials to conditions. *Behavior Research Methods*, *47*, 127–133. doi:10.3758/s13428-014-0476-9

Zeuch, N., Holling, H., & Kuhn, J.-T. (2011). Analysis of the Latin Square Task with linear logistic test models. *Learning and Individual Differences*, *21*, 629–632. doi:10.1016/j.lindif.2011.03.004

Footnotes

[1]Some authors cite this work as “Ozanam (1725)”. At the time, it was common to show the date of printing on the title page of a book rather than its date of publication. The authors in question appear to have obtained reprints of this work from 1725.

[2]In both the abstract and the text of the article by Willner-Reid et al. (2016), it is stated that a 2×2 Latin-square design was used. The independent variables in their study did indeed define a 2×2 repeated-measures design, but the order of administration of the four resulting conditions was counterbalanced by means of a 4×4 Latin square (K. L. Preston, personal communication, January 26, 2017).

Table 1

Fisher's (1925) analysis of data from a 5×5 Latin square.

Source of variation	Sum of squares	Degrees of freedom	Mean square
Rows	4240.24	4	1060.06
Columns	701.84	4	175.46
Treatments	330.24	4	82.56
Residual	1754.32	12	146.19
Total	7026.64	24	292.78

Note: Adapted from Fisher (1925, p. 230).

Table 2

Nisbet's (1939) research design.

	Test 1	Test 2	Test 3	Test 4
Group I	List A	List B	List C	List D
Group II	List B	List C	List D	List A
Group III	List C	List D	List A	List B
Group IV	List D	List A	List B	List C

Note: Adapted from Nisbet (1939, p. 34).

Table 3

The research design described by Garrett and Zubin (1943).

	Illumination level			
Group	1	2	3	4
1	Red	Blue	Yellow	Green
2	Green	Red	Blue	Yellow
3	Blue	Yellow	Green	Red
4	Yellow	Green	Red	Blue

Note: Adapted from Garrett and Zubin (1943, p. 243).

Table 4

A 3 × 3 Latin-square design assigning three prose passages to three instructional sets for three groups of participants.

	Instructional set		
Group	1	2	3
1	Passage 2	Passage 1	Passage 3
2	Passage 3	Passage 2	Passage 1
3	Passage 1	Passage3	Passage 2

Table 5

Hypothetical data from an experiment on the recall of prose passages.

		Instructional set		
Group	Participant	1	2	3
1	1	9	8	15
1	2	7	9	14
2	3	12	12	7
2	4	14	14	5
3	5	3	19	11
3	6	3	18	10

Table 6

Summary table from analysis of variance ignoring the Latin-square design in Table 5.

Source	Sum of squares	Degrees of freedom	Mean square	*F*
Between subjects	2.444	5	0.489	
Within subjects	366.000	12	30.500	
Instructional set	85.778	2	42.889	1.53
Residual	280.222	10	28.022	

Table 7

Summary table from analysis of variance incorporating the Latin-square design in Table 5.

Source	Sum of squares	Degrees of freedom	Mean square	*F*
Between subjects	2.444	5	0.489	
Groups	0.444	2	0.222	0.33
Residual	2.000	3	0.667	
Within subjects	366.000	12	30.500	
Instructional set	85.778	2	42.889	32.17***
Passage	270.778	2	135.389	101.54***
Set × Passage	1.444	2	0.722	0.54
Residual	8.000	6	1.333	

***$p < .001$.

Table 8

The row-complete Latin-square design described by Blumberg et al. (1983).

	Order					
Group	1st	2nd	3rd	4th	5th	6th
1	5	2	3	1	6	4
2	2	1	5	4	3	6
3	3	5	6	2	4	1
4	1	4	2	6	5	3
5	6	3	4	5	1	2
6	4	6	1	3	2	5

Note: The figures in the cells show the order of administration of the six different texts (shown as 1–6). Each text is followed only once by each of the other five texts.

Figure Captions

Fig. 1. A 4×4 Latin square.

Fig. 2. Ozanam's (1723) solution to the playing-card problem. Playing cards by Trocche100 at Italian Wikipedia, transferred from it.wikipedia to Commons, Public Domain, https://commons.wikimedia.org/w/index.php?curid=37086025

Fig. 3. The first 4×4 Latin square in Ozanam's (1723) solution.

Fig. 4. The 2×2 Latin squares.

Figure 1

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

Figure 2

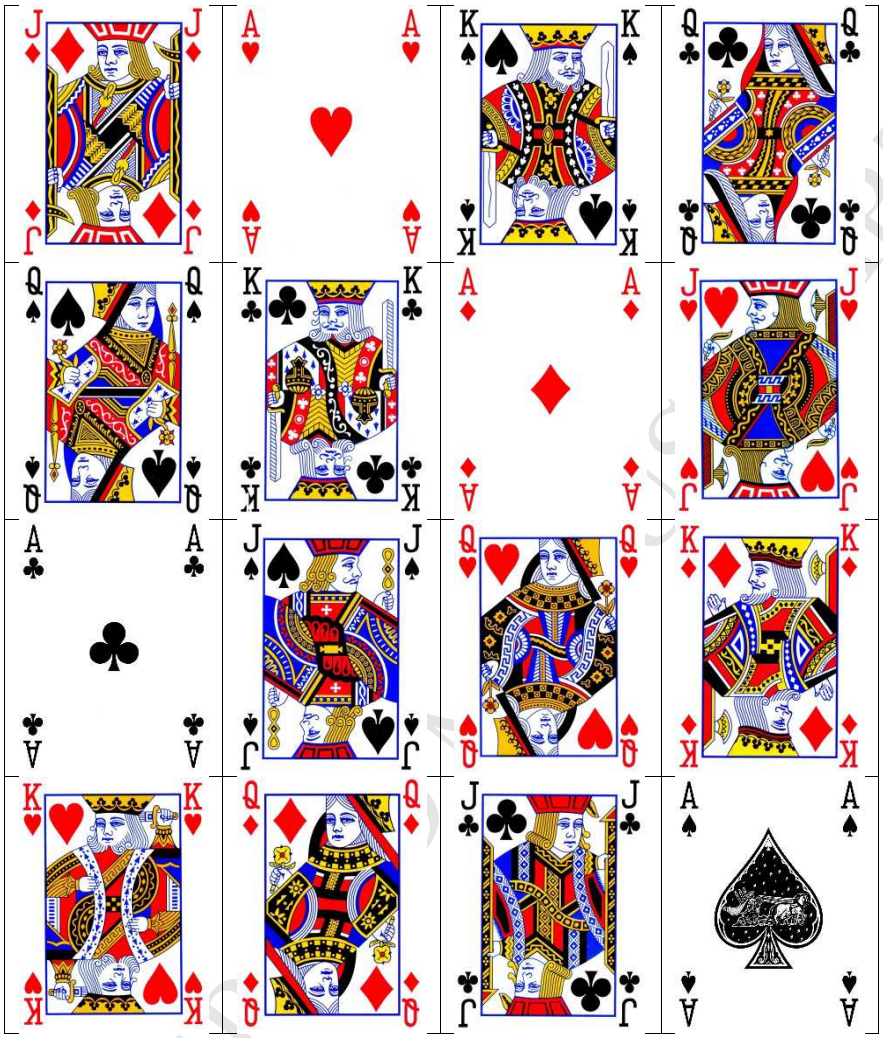
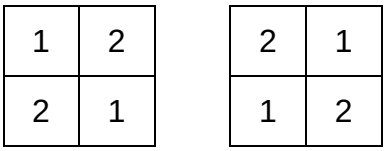


Figure 3

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

Figure 4



Highlights

- Fisher (1925) proposed that Latin squares could be useful in experimental design.
- He argued that Latin squares should be chosen at random and used in data analysis.
- Educational and psychological researchers have used Latin squares only rarely.
- Those who have used Latin squares have often not heeded Fisher's prescriptions.
- Nevertheless, the judicious use of Latin-square designs can be a powerful tool.